Machine Learning Approaches to Analyzing Path Loss and Throughput in 5G Networks

Md. Zahidul Islam
New York University
Email: mi2502@nyu.edu

Asad Anjum
New York University
Email: aa11227@nyu.edu

***Abstract*—This study involves testing and predicting 5G coverage in diverse environments using multiple datasets. These datasets encompass signal strength, quality, angle, and distance metrics of 5G signals in environments ranging from open rural spaces to densely packed urban areas. Our approach involved selecting critical features from the datasets, training a machine learning model, and utilizing the model to predict 5G coverage outcomes. A wide range of connections between features and outcomes was discovered, and multiple machine learning methods were tested to determine which produced the most accurate predictions.**

I. Introduction

5G technology has brought transformative potential to wireless networking by significantly enhancing speed, reducing latency, and increasing connectivity. However, the deployment and optimization of 5G networks in different environments has presented unique challenges, leading to an influx of research around path loss and throughput analysis. Work from Kohli et al [1] monitored 5G coverage in indoor and outdoor settings, citing the impact of different materials on signal quality, while Yuliana et al [2] studied the effect of altering a litany of different physical and environmental features, testing multiple machine learning models in the process to see which algorithms would best predict coverage.

This study focuses on understanding and predicting 5G coverage using machine learning techniques across different environments, utilizing datasets that capture a wide range of 5G signal characteristics. By analyzing these datasets, which include both path loss and throughput metrics under various conditions, we aim to develop robust models that can accurately forecast 5G network performance and coverage. This understanding is critical for optimizing 5G deployment strategies in diverse geographical and infrastructural landscapes, enhancing user experience and network reliability.

A multitude of different machine learning algorithms were used to make predictions for 5G coverage based on a preordained set of features, and their accuracy's were compared in order to find the most effective model. Due to the unfortunate lack of certain key information in many datasets online, the feature sets could not be as comprehensive as initially intended, however the findings still suggest that 5G data coverage can be accurately predicted in any setting given a few important metrics.

II. Description of the Datasets

The datasets have been taken from publicly available 5G experiments which capture signal data in a variety of different environments. Models were trained separately for each environment, meaning the prediction for coverage in a sparsely populated area would differ from a dense, downtown region. Our study utilized two primary datasets to analyze and predict 5G coverage: the Path Loss Dataset [3] and the Throughput Dataset [4], each providing unique insights into 5G signal behavior in different environments.

A. Path Loss Dataset

This dataset includes measurements like elevation angle, azimuth angle, line of sight conditions, and Signal-to-Interference-plus-Noise Ratio (SINR), segmented by rural, suburban, and urban settings. It provides a structured analysis of how signal path loss varies with geographical and infrastructural changes. A limitation is the absence of transmitted power levels, which are critical for accurate path loss calculations. The original paper did note a maximum transmitted power level however, therefore for our analysis we assumed the tower was operating at maximum power for the duration of the testing. The initial study applied empirical methods with established path loss models to this data, refining 5G signal predictions by adjusting model parameters to better fit the measured environmental conditions. It also required preprocessing to eliminate NaN values, calculate distance based off longitude and latitude values, and truncating the data to only list features which we found relevant to our own study. The features we decided to train our models on were:

- **Elevation Angle:** The angle between the transmitter and receiver, affecting the signal's trajectory.
- **Azimuth Angle:** The compass direction from the receiver to the transmitter, influencing directional signal propagation.
- **Line of Sight Conditions:** Indicates whether the signal path is obstructed, which critically affects signal strength and quality.
- **SINR (Signal-to-Interference-plus-Noise Ratio):** A measure of signal quality relative to the background interference and noise.

The relevant features used for the path loss study are shown in Table I.

TABLE I: Path Loss Dataset Snippet

Distance	Near	Azimuth	Elevation	RSRP	SINR	TX Power	Path Loss
1.30	1	-48.75	10	-79	33.18	61.93	140.93
1.30	1	-56.25	10	-79	32.36	61.93	140.93
1.30	1	-56.25	10	-82	32.13	61.93	143.93
1.30	1	-48.75	10	-80	33.33	61.93	141.93
1.30	1	-56.25	10	-81	33.49	61.93	142.93
1.31	1	-56.25	10	-82	33.61	61.93	143.93
1.33	1	-48.75	10	-79	34.49	61.93	140.93
1.35	1	-48.75	10	-79	34.17	61.93	140.93

B. Throughput Dataset

This dataset captures end device throughput performance metrics like RSSI, RSRP, RSRQ, and receiver speed, but lacks precise distance measurements between transmitters and receivers, only including receiver coordinates. Original analyses employed statistical techniques to correlate throughput with signal quality and mobility, using regression to predict throughput based on these factors. The throughput was noted as being at a high (¿700 Mbps), medium (¿300 Mbps, ¡700 Mbps), or low (¡300 Mbps) level. Like the previous dataset, this also required pre-processing, including separating the data into 3 skews representing each classification for throughput so that it could be analyzed as a classification problem instead. The features we decided to train our models on were:

- **Received Signal Strength Indicator (RSSI):** A measure of the power level that a device is receiving from the wireless network.
- **Reference Signal Received Power (RSRP):** The power of the LTE reference signals spread over the full bandwidth and is a more accurate depiction of signal strength.
- **Reference Signal Received Quality (RSRQ):** A ratio of RSRP to the total received signal power, indicating cell loading.
- **Receiver Speed:** The speed at which the receiving device is moving, which can affect the signal dynamics due to the Doppler effect.

The relevant features used for the throughput study are shown in Table II.

TABLE II: Throughput Dataset Snippet

Speed	NR Status	RSSI	RSRP	RSRQ	RSSNR	Throughput
7.48	CONNECTED	-59.0	-95	-16.0	2.15e+09	4
7.51	CONNECTED	-59.0	-95	-16.0	2.15e+09	4
7.64	CONNECTED	-59.0	-95	-16.0	2.15e+09	116
8.97	CONNECTED	-51.0	-81	-10.0	2.15e+09	112
8.99	CONNECTED	-51.0	-81	-10.0	2.15e+09	54
8.93	CONNECTED	-51.0	-81	-10.0	2.15e+09	13
8.71	CONNECTED	-51.0	-81	-10.0	2.15e+09	106
7.17	CONNECTED	-51.0	-81	-10.0	2.15e+09	121
6.45	CONNECTED	-51.0	-81	-10.0	2.15e+09	96
5.97	CONNECTED	-51.0	-81	-10.0	2.15e+09	101

Our methodology enhances these analyses by incorporating different machine learning models to dynamically predict 5G coverage, and considering the complex interplay between environmental factors, signal characteristics, and user mobility. This streamlined approach maintains the focus on environment and mobility segmentation from earlier studies while basing predictions on only a short list of features.

III. Methodologies

Machine learning (ML) models are widely employed in various fields for data analysis and prediction tasks. Two fundamental types of ML models are regression and classification. In this context, path loss analysis is treated as a regression task, whereas throughput analysis is approached as a classification task.

Regression models are used to predict continuous numerical values based on input features. In the context of path loss estimation, regression models can be employed to predict the path loss between a transmitter and receiver based on factors such as distance, frequency, and environmental conditions. One commonly used regression model is the linear regression model, which predicts the target variable as a linear combination of the input features. The model can be represented by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

where y is the predicted path loss, $x_1, x_2, \ldots, x_n$ are the input features, $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients, and ϵ is the error term.

Classification models are used to categorize data into different classes or categories. In the context of throughput analysis, classification models can be employed to classify the network states into categories such as high throughput, medium throughput, and low throughput based on factors such as signal strength, interference, and modulation scheme. One commonly used classification model is the logistic regression model, which predicts the probability of an observation belonging to a particular class. The model can be represented by the equation:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

where $P(Y = 1|X)$ is the probability of the observation belonging to class 1, X is the input feature vector, and $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients.

These regression and classification models play a crucial role in estimating path loss and analyzing throughput in communication systems, providing valuable insights for network optimization and performance improvement. For our analysis, we use five different models for the regression and classification purposes.

In this section, we provide a brief overview of the machine learning methods employed for both regression and classification tasks, including Random Forest (RF), XGBoost (XGB), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN).

A. Random Forest (RF)

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mean prediction of the individual trees for regression tasks or the mode prediction for classification tasks. In regression, the final prediction of the Random Forest model is the average of predictions from all trees, while in classification, it's the majority vote of predictions. The method is known for its robustness, scalability, and ability to handle high-dimensional data.

B. XGBoost (XGB)

XGBoost is a powerful gradient boosting algorithm that iteratively builds an ensemble of weak learners, typically decision trees, to minimize a loss function and improve predictive performance. It excels in both regression and classification tasks by optimizing a differentiable loss function, such as mean squared error for regression and cross-entropy for classification. XGBoost is highly efficient, scalable, and offers excellent accuracy, making it a popular choice for various machine learning applications.

C. Multilayer Perceptron (MLP)

Multilayer Perceptron is a type of artificial neural network that consists of multiple layers of interconnected neurons, including an input layer, one or more hidden layers, and an output layer. MLPs are trained using backpropagation and gradient descent algorithms to minimize the error between predicted and actual values. They are versatile and can be applied to both regression and classification tasks, with the output layer tailored to the specific problem.

D. Support Vector Machine (SVM)

Support Vector Machine is a supervised learning algorithm that constructs a hyperplane or set of hyperplanes in a high-dimensional space to separate data points into different classes for classification or predict continuous outcomes for regression. SVMs aim to maximize the margin between classes while minimizing classification errors. They are effective for both linear and nonlinear problems and can handle high-dimensional data efficiently.

E. k-Nearest Neighbors (KNN)

k-Nearest Neighbors is a non-parametric algorithm that makes predictions based on the majority class or average of the k-nearest data points in the feature space. In regression, KNN calculates the average of the target values of the k-nearest neighbors, while in classification, it assigns the class label that is most common among the k-nearest neighbors. KNN is simple, intuitive, and suitable for both regression and classification tasks.

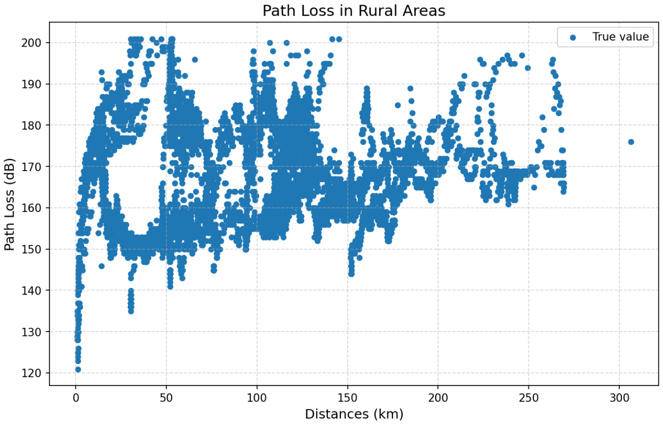

Fig. 1: Actual path loss in rural area

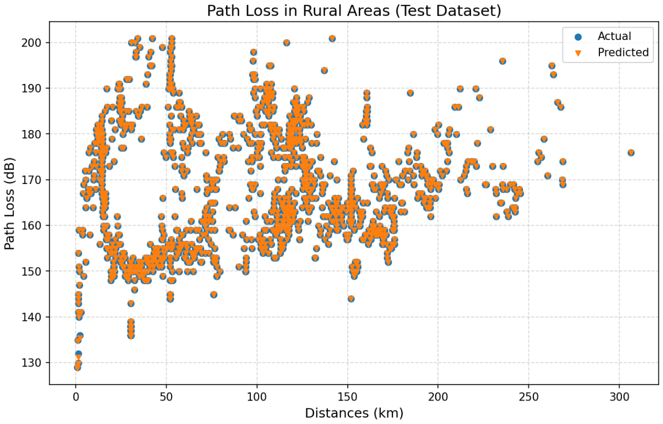

Fig. 2: Actual Vs. predicted path loss in rural area using RF model

These machine learning methods offer a wide range of capabilities and can be effectively applied to regression (i.e., path loss estimation) and classification (i.e., throughput classes) tasks in wireless networks, providing valuable insights and predictive performance.

IV. Results

For the regression task, we utilize path loss dataset to predict the path losses in different areas. The path loss dataset comprises three distinct subsets corresponding to rural, suburban, and urban areas. Each subset was analyzed using ML models available in both the scikit-learn and XGBoost libraries in Python. To apply ML models to the path loss dataset, we divided the dataset into training and testing sets using a ratio of 8:2 and employed the shuffle option provided by the *train_test_split* function in Python's scikit-learn library. After training the ML models on the training dataset, we validated the predictions against the test dataset using three error metrics: mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). Figures 1, 2, and 3 illustrate the ground truth, predictions on the test set, and prediction errors, respectively, for the rural

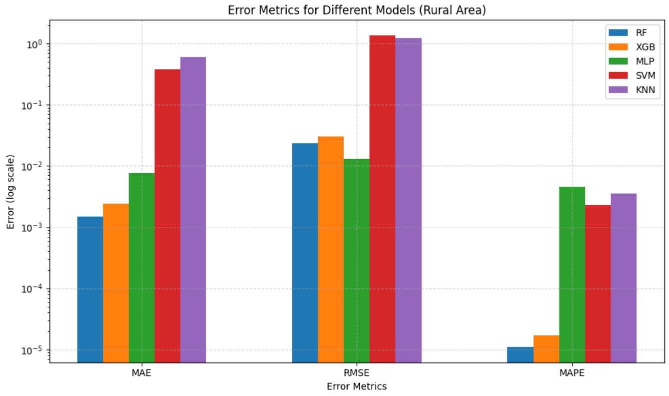

Fig. 3: Error metrics for different ML models in rural area dataset.

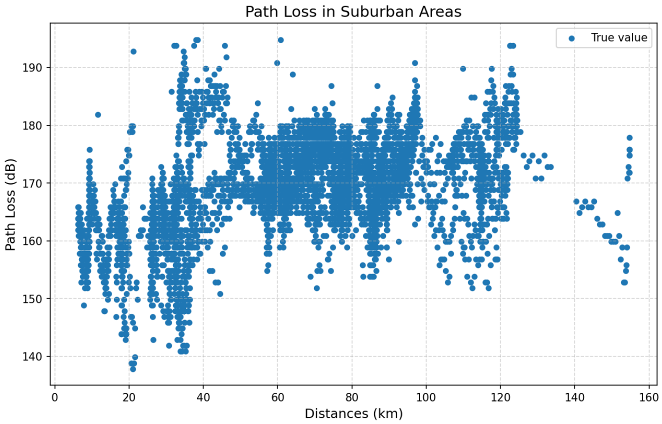

Fig. 4: Actual path loss in suburban area

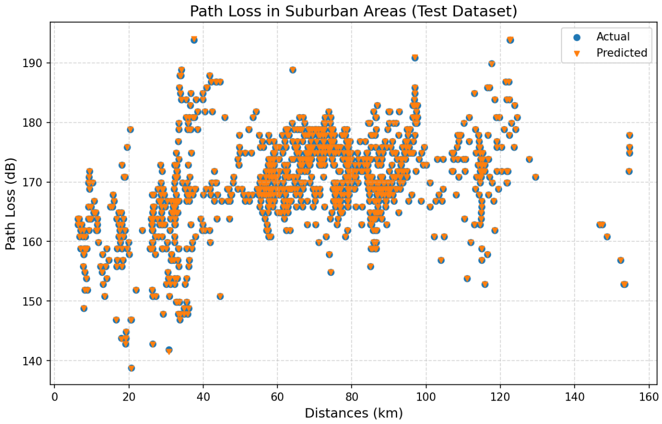

Fig. 5: Actual Vs. predicted path loss in suburban area using RF model

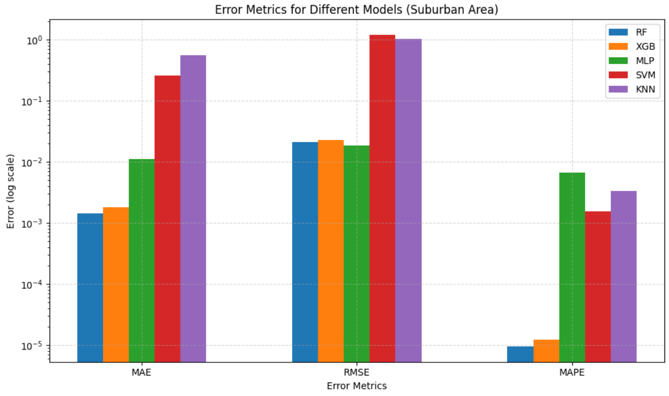

Fig. 6: Error metrics for different ML models in suburban area dataset.

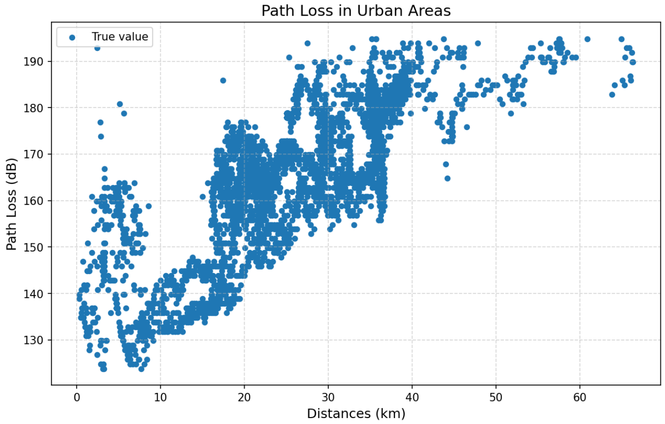

Fig. 7: Actual path loss in urban area

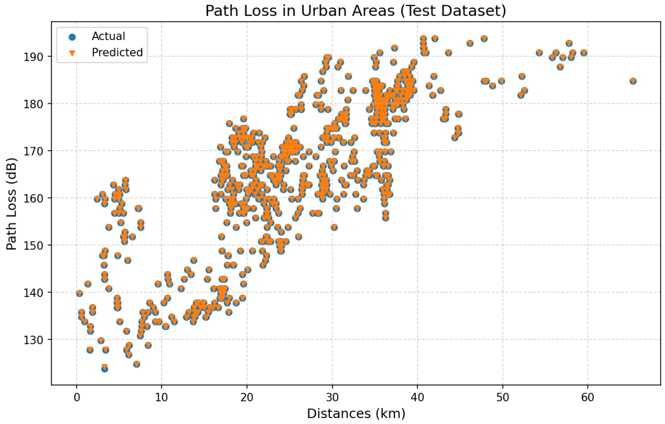

Fig. 8: Actual Vs. predicted path loss in urban area using RF model

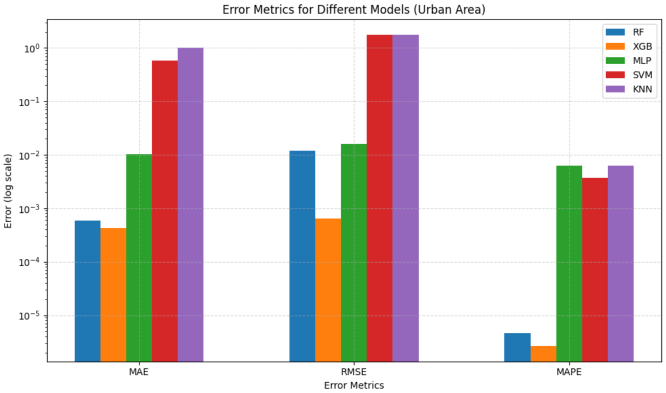

Fig. 9: Error metrics for different ML models in urban area dataset.

path loss dataset. Similar sets of figures are provided for the suburban and urban path loss datasets, as shown in figures 4-6, and 7-9. Our analysis indicates that the Random Forest (RF) and XGBoost (XGB) models generally yield more accurate estimations compared to other models, as evidenced by both visual inspection of the figures and evaluation using the error metrics.

For the classfication task, we utilize the throughput dataset to classify throughput measurements into three categories: low, medium, and high. The thresholds for these categories are defined following the ranges specified in [4]. Upon training the machine learning models, we assess their performance using two principal metrics: accuracy and the F1 score. **Accuracy** measures the proportion of true results among the total number of cases examined. It is defined as the ratio of correctly predicted observations (both true positives and true negatives) to the total observations:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP represents true positives, TN true negatives, FP false positives, and FN false negatives. **F1 Score**, on the other hand, merges precision and recall into a singular measure, making it particularly valuable in scenarios where class distribution is uneven and one class may be significantly underrepresented. The F1 score is the harmonic mean of precision and recall:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where *Precision* is $\frac{TP}{TP+FP}$ and *Recall* is $\frac{TP}{TP+FN}$.

Performance metrics for the machine learning models are detailed in Table III. The results indicate that the Random Forest (RF) model outperforms other models. Figure 10 presents the confusion matrix for the RF model, highlighting its overall performance. It is noted that there is a lower degree of confusion for the medium throughput label, which can be attributed to the limited number of training samples available for this category. The overall analysis demonstrates the adaptability of the machine learning models in analyzing throughput, effectively managing variations in data distribution and sample availability.

TABLE III: Performance Metrics for Throughput Classification.

Metric	RF	XGB	MLP	SVM	KNN
Accuracy	0.82	0.76	0.76	0.62	0.73
F1-score	0.81	0.75	0.75	0.57	0.73

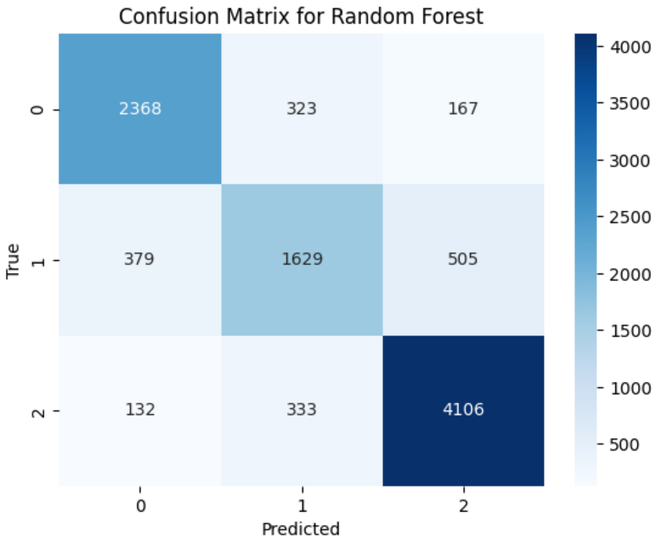

Fig. 10: Confusion metric for RF model in classifying the throughput dataset.

V. Conclusion

In this report, we have analyzed two distinct, publicly available datasets to examine path loss across different environments and to determine the achievable throughput in various scenarios within 5G networks. Specifically, our path loss analysis forecasts the loss in rural, suburban, and urban settings, while the throughput analysis classifies the levels of throughput achievable under certain signal conditions.

To achieve these objectives, we employed a range of machine learning models to evaluate their effectiveness in both path loss and throughput prediction tasks. The simulation results indicate that the Random Forest model consistently outperforms other models in both domains. This finding underscores the robustness of the Random Forest algorithm in handling the complexities associated with 5G network parameters.

This comprehensive analysis offers significant insights into data-driven parameter modeling within 5G networks, which are critical for the planning and optimization tasks. Such insights can significantly enhance planning and optimization strategies, thereby facilitating more efficient network design and improved service delivery in 5G deployments.

References

[1] Manav Kohli, Abhishek Adhikari, Gulnur Avci, Sienna Brent, Jared Moser, Sabbir Hossain, Aditya Dash, Igor Kadota, Rodolfo Feick, Dmitry Chizhik, Jinfeng Du, Reinaldo A. Valenzuela, and Gil Zussman1. Outdoor-to-indoor 28 ghz wireless measurements in manhattan: Path loss, location impacts, and 90% coverage. In *Proceedings of the Association for Computing Machinery*. ACM, 2022.

[2] Hajiar Yuliana, Iskandar, and Hendrawan. Comparative analysis of machine learning algorithms for 5g coverage prediction: Identification of dominant feature parameters and prediction accuracy. *IEEE Access*, 12:18939–18956, 2024.

[3] Adrian Schumacher, Ruben Merz, and Andreas Burg. 3.5 ghz coverage assessment with a 5g testbed. In *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*, pages 1–6, 2019.

[4] Arvind Narayanan, Eman Ramadan, Rishabh Mehta, Xinyue Hu, Qingxu Liu, Rostand AK Fezeu, Udhaya Kumar Dayalan, Saurabh Verma, Peiqi Ji, Tao Li, et al. Lumos5g: Mapping and predicting commercial mmwave 5g throughput. In *Proceedings of the ACM Internet Measurement Conference*, pages 176–193, 2020.